The Variational Bayesian Approach to Fitting Mixture Models to Circular Wave Direction Data

BURTON WU

Queensland University of Technology, Brisbane, Queensland, Australia

CLARE A. MCGRORY

The University of Queensland, Brisbane, Queensland, Australia

ANTHONY N. PETTITT

Queensland University of Technology, Brisbane, Queensland, Australia

(Manuscript received 20 June 2011, in final form 5 July 2012)

ABSTRACT

The emerging variational Bayesian (VB) technique for approximate Bayesian statistical inference is a non-simulation-based and time-efficient approach. It provides a useful, practical alternative to other Bayesian statistical approaches such as Markov chain Monte Carlo–based techniques, particularly for applications involving large datasets. This article reviews the increasingly popular VB statistical approach and illustrates how it can be used to fit Gaussian mixture models to circular wave direction data. This is done by taking the straightforward approach of padding the data; this method involves adding a repeat of a complete cycle of the data to the existing dataset to obtain a dataset on the real line. The padded dataset can then be analyzed using the standard VB technique. This results in a practical, efficient approach that is also appropriate for modeling other types of circular, or directional, data such as wind direction.

1. Introduction

Mixture models provide a convenient, flexible way to model data; a popular and often appropriate approach is to fit a Gaussian mixture model (GMM; McLachlan and Peel 2000). A mixture model provides a way to represent a complicated density as a linear combination of simpler densities, which are called the mixture components. The parameters of these components are estimated as part of a statistical analysis. In this paper, we describe the key ideas of the increasingly popular variational Bayesian (VB) method for Bayesian statistical inference; VB has been shown to approximate Bayesian posterior distributions efficiently, and, in particular, it has been shown to be very useful for fitting mixture models (McGrory and Titterington 2007). This paper follows and reviews the standard VB approach for modeling one-dimensional data with GMMs that was described in McGrory and Titterington (2007) (note that throughout we refer to this as the VB-GMM algorithm). In addition, this article makes the contribution of describing how the standard VB approach can straightforwardly be applied to the circular data problem of modeling wave directions by using a data-padding approach. The ideas described here are also more generally applicable to any circular data problem—for example, modeling wind direction.

Our application involves analyzing hindcast wave direction data; we focus on modeling daily mean wave directions off the coast of Byron Bay in southeastern Australia, over a period of 45 yr. The daily mean wave directions are periodic observations mapped onto the circle that take values between 0 and 2π. It is important to take the circular characteristics of the wave direction data into consideration when analyzing it. To see why this is important, consider for instance the distance between an observation at 0 and one just before 2π, on the circle these observations are close to one another whereas on the real line they are not (e.g., see Farrugia et al. 2009;

Corresponding author address: Clare A. McGrory, Centre for Applications in Natural Resource Mathematics, School of Mathematics, The University of Queensland, St. Lucia, Brisbane, QLD 4072, Australia.
E-mail: c.mcgrory@uq.edu.au

DOI: 10.1175/JAMC-D-11-0124.1

Mahrt 2011; Weber 1997). This must be accounted for when carrying out a statistical analysis of this type of data. Data analysis of this kind is often referred to as circular or directional statistics (Mardia and Jupp 2000; Jammalamadaka and Sengupta 2001). There have been numerous distributional models already proposed for analyzing circular data, and the von Mises method, also known as circular normal (Jammalamadaka and Sengupta 2001), is perhaps the most popular choice (McVinish and Mengersen 2008). Fitting this type of model can be very computationally demanding, however. Another option is to use nonparametric kernel density estimation based on the von Mises–Fisher kernel (Mardia and Jupp 2000, 277–278), but the results of the kernel approach depend on the degree of smoothing and lack the interpretability that may be critical for some applications. In this paper, we propose tackling the circular data problem by using the more straightforward intuitive approach of padding the repeated data at both ends (see Mardia and Jupp 2000, p. 4) and then normalizing the resulting models [i.e., $f(x; 0 \leq x < 2\pi)$]. In this way we create a padded version of the original dataset that can be analyzed using standard methods for linear data analysis. This means that we can then apply the VB-GMM approach to the padded dataset; VB is an approximate approach to Bayesian inference for this type of problem but still allows us to obtain a very good fit to the data in a much more time-efficient analysis than would be possible using other more popular Bayesian approaches.

The VB approach to Bayesian inference is more efficient in terms of both computation and storage requirements than most other Bayesian statistical approaches such as the reversible jump Markov chain Monte Carlo technique (RJMCMC; Richardson and Green 1997). In addition, unlike Monte Carlo–based approaches, VB does not suffer from the label-switching problem when fitting mixture models (see Celeux et al. 2000) or from the difficulties with assessing convergence (McGrory and Titterington 2007). In addition, because it is a Bayesian method, VB suffers less from the overfitting and singularity problems that persist in maximum likelihood (ML) approaches (Attias 1999). Given that a central issue of mixture modeling is the selection of a suitable number of components k (McLachlan and Peel 2000), a key practical advantage of VB over ML approaches is its ability to automatically select k to give the ''best'' fit to the data according to the variational approximation and to estimate the model parameter values and their posterior distributions at the same time. Standard VB-based algorithms achieve this through the complexity reduction property of the VB approximation; this property leads to the progressive elimination of redundant components that were specified in the initial model during convergence (McGrory and Titterington 2007). Note that this implies that the final k, k_{final}, in the model cannot be greater than the initial specification of k, k_{initial}.

The useful automatic feature of the approximation has been observed by many researchers (e.g., Attias 1999; Corduneanu and Bishop 2001; McGrory and Titterington 2007). Its theoretical reasoning is still not well understood, however. Therefore, we point out that this attribute might be viewed as a drawback by some researchers instead of an advantage. When using the standard VB algorithm, there is the possibility that sometimes the selected number of components k in the final model might vary depending on how the scheme is initialized. This is one limitation of using the standard approach, but it is part of the nature of mixture modeling that different fitted models with different values of k can provide good representations of the same dataset. More-advanced VB schemes have been proposed, and, although we did not choose to do so in this paper, these could be explored for this type of application. Such schemes involve component splitting that allows the number of mixture components to be increased as well as decreased during the convergence of the VB algorithm, thereby providing increased flexibility [see Wu et al. (2012) and references therein for further reading on component-splitting VB schemes].

In section 2 we describe the VB-GMM algorithm and the data padding. In section 3 we describe our application dataset, and in section 4 we present the results of our analysis of it. Section 5 concludes the article.

2. VB-GMM algorithm

In a GMM, it is assumed that all k underlying distributions (or components) of the mixture are Gaussian. In the notation we adopt here, the mixture model density of an observation $x = (x_1, \ldots, x_n)$ on the real line is then given by $\sum_{j=1}^{k} w_j N(x; \mu_j, \tau_j^{-1})$, where $N()$ denotes a Gaussian density, k is the number of components, and μ_j and τ_j^{-1} correspond to the mean and variance, respectively, of the jth component. Each mixing coefficient w_j satisfies $0 \leq w_j$ and $\sum_{j=1}^{k} w_j = 1$. In the Bayesian framework, inference is based on the target posterior distribution, $p(\theta, z \mid x)$, where θ denotes the model parameters (μ, τ, w) and $z = \{z_{ij}\}$ denotes the missing component membership information of observation x. Note that the z_{ij}s are indicator variables such that $z_{ij} = 1$ if observation x_i belongs to the jth component and $z_{ij} = 0$ otherwise.

The target posterior is not analytically available in this mixture model problem, as is generally the case, and therefore it has to be estimated in the Bayesian inference approach. The idea of the VB approach is to approximate the target posterior by a variational distribution that we

denote by $q(\theta, z|x)$. It is assumed that this approximating distribution factorizes over the model parameters θ and the missing variables z; this assumption means that we can write $q(\theta, z|x) = q_\theta(\theta|x) \times q(z|x)$. To obtain a good approximation to the target, the distribution $q(\theta, z|x)$ is chosen to maximize the lower bound on the log marginal likelihood. Note that this is equivalent to minimizing the Kullback–Leibler (KL) divergence between the target posterior and the variational approximating distribution. This approach leads to tractable coupled expressions for the variational posterior over the parameters that can be iteratively updated to obtain convergence to a solution. Convergence to at least a local minimum is guaranteed. Carefully choosing the initialization settings for the algorithm makes it likely that the minimum that is found, if not the global minimum, is sufficiently close to the global minimum. In terms of how this compares with popular alternative approaches for fitting mixtures, we note that it is also the case with the well-known and popular classical expectation–maximization approach that only local convergence is guaranteed. Also, whereas if the algorithm is run for long enough then in theory a Bayesian Markov chain Monte Carlo (MCMC) approach should fully explore the posterior and converge to a global maximum, in practice chains can become stuck in local minima, and it can be difficult to assess just how long of a run is required to reach convergence.

In the machine-learning literature, VB has been used for performing approximate Bayesian inference since the late 1990s, and we refer readers to Mackay (2003) and Bishop (2006) for more background on the approach. Most of the papers on the subject of fitting GMMs with VB (e.g., Attias 1999; Corduneanu and Bishop 2001; McGrory and Titterington 2007) make similar prior assumptions, but they differ in the form of the model hierarchy used. As indicated previously, we follow the model setting and algorithm described in McGrory and Titterington (2007). Note that to apply this algorithm to our circular data we have to first pad the data, as we will describe later.

a. *The standard VB-GMM algorithm (McGrory and Titterington 2007)*

We model the pattern as a mixture of k Gaussian distributions with unknown means $\mu = (\mu_1, \ldots, \mu_k)$, precisions $\tau = (\tau_1, \ldots, \tau_k)$, and mixing coefficients $w = (w_1, \ldots, w_k)$, such that

$$p(x, z \mid \theta) = \prod_{i=1}^{n} \prod_{j=1}^{k} [w_j N(x_i; \mu_j, \tau_j^{-1})]^{z_{ij}},$$

with the joint distribution being $p(x, z, \theta) = p(x, z|\theta)\, p(w)p(\mu|\tau)p(\tau)$. We express our priors as

$$p(w) = \text{Dirichlet}[w; \alpha_1^{(0)}, \ldots, \alpha_k^{(0)}],$$

$$p(\mu \mid \tau) = \prod_{j=1}^{k} N\{\mu_j; m_j^{(0)}, [\beta_j^{(0)}\tau_j]^{-1}\}, \quad \text{and}$$

$$p(\tau) = \prod_{j=1}^{k} \text{gamma}\left[\tau_j; \frac{1}{2}\upsilon_j^{(0)}, \frac{1}{2}\sigma_j^{(0)}\right],$$

with $\alpha^{(0)}$, $\beta^{(0)}$, $m^{(0)}$, $\upsilon^{(0)}$, and $\sigma^{(0)}$ being the hyperparameter values, which are chosen by the user. These are the standard conjugate priors used in Bayesian mixture modeling (Gelman et al. 2003). Using the lower bound approximation, the posteriors are then

$$q_w(w) = \text{Dirichlet}(w; \alpha_1, \ldots, \alpha_k),$$

$$q_{\mu|\tau}(\mu \mid \tau) = \prod_{j=1}^{k} N[\mu_j; m_j, (\beta_j\tau_j)^{-1}], \quad \text{and}$$

$$q_\tau(\tau) = \prod_{j=1}^{k} \text{gamma}\left(\tau_j; \frac{1}{2}\upsilon_j, \frac{1}{2}\sigma_j\right).$$

The posterior parameters are iteratively updated as [see McGrory and Titterington (2007) for further details]

$$\alpha_j = \alpha_j^{(0)} + \sum_{i=1}^{n} q_{ij}, \quad \beta_j = \beta_j^{(0)} + \sum_{i=1}^{n} q_{ij},$$

$$\upsilon_j = \upsilon_j^{(0)} + \sum_{i=1}^{n} q_{ij}, \quad m_j = \frac{1}{\beta_j}\left[\beta_j^{(0)} m_j^{(0)} + \sum_{i=1}^{n} q_{ij} x_i\right], \quad \text{and}$$

$$\sigma_j = \sigma_j^{(0)} + \sum_{i=1}^{n} q_{ij} x_i^2 + \beta_j^{(0)} m_j^{(0)^2} - \beta_j m_j^2,$$

where expectations are given by $E(\mu_j) = m_j$, and $E(\tau_j) = \upsilon_j \sigma_j^{-1}$. Note that q_{ij} is the VB posterior probability that $z_{ij} = 1$, and the update expression for this quantity is given by

$$q_{ij} = \frac{\exp\left\{\Psi(\alpha_j) - \Psi\left(\sum_j \alpha_j\right) + \frac{1}{2}\left[\Psi\left(\frac{1}{2}\gamma_j\right) - \log\frac{\delta_j}{2}\right] - \frac{1}{2\beta_j} - \frac{\gamma_j}{2\delta_j}(x_i - m_j)^2\right\}}{g_i},$$

where Ψ is the digamma function and g_i is a normalizing constant. This expression is normalized so that for each observation x_i, the q_{ij}s sum to 1 over the js.

We can see that the updates for the q_{ij}s and for the posterior parameters are a set of coupled expressions (i.e., they each involve one another); therefore, they must be solved iteratively. In other words, to implement this approach, the user must choose some initial values for the number of components k, the hyperparameters, and the q_{ij}s and then proceed to alternatively update the expressions given above for the q_{ij}s and the parameters. The $\sum_{i=1}^{n} q_{ij}$ for each component $j = 1, \ldots, k$ corresponds to the estimated number of observations that belong to that component. As the expressions are iteratively updated, if the estimated number of observations for any component drops below 1, then that component can be removed from the algorithm. In this way, once convergence is reached, only components that are estimated as being useful in the model will remain. Once this alternative updating of the expressions no longer changes the values, the algorithm can be declared to have converged to give the required parameter estimates for the fitted mixture model and the number of components with their corresponding weights.

b. *Padding the circular data to apply the standard VB-GMM algorithm*

In the approach we have just outlined, the observations $x = (x_1, \ldots, x_n)$ are assumed to be on the real line. If we have data that are measured on the circle, we have to adjust them before we can use it. As mentioned, we will take the straightforward approach of padding the data by adding a repeat of a complete cycle of the data to the existing dataset to obtain a dataset on the real line (Mardia and Jupp 2000). Without padding, the data may have modes located around 0 (or 2π) that would create a problem when trying to model the data with an approach that assumes the data are on the real line. We pad the data as follows. For each observed value x_i, where $i = \{1, \ldots, n\}$,

$$\begin{aligned} \text{if observation} \quad & x_i < \pi \\ & p_i = x_i + 2\pi, \\ \text{otherwise} \quad & p_i = x_i - 2\pi. \end{aligned}$$

This new set of padded observations $\{p_i\}, i = 1, \ldots, n$, corresponds to a complete repeated cycle of the original data around the circle. The original data and the set of padded data are then combined to produce the dataset used in the VB-GMM analysis.

3. Wave direction data

Our application involves analyzing wave direction data; in particular, we focus on wave directions off the coast of Byron Bay in New South Wales (NSW), Australia. There is significant interest in monitoring of waves along the NSW coast because of their potentially damaging impact on the coastline (e.g., Shand et al. 2010). At Byron Bay, and in southeastern Australia generally, waves typically propagate toward the coast from the east-southeast to south (Goodwin 2005). Goodwin (2005) explains that this wave climate creates a longshore sand transport system; therefore, the coastline stability in southeastern Australia is closely related to temporal changes in sand transport brought about by changes in wave-driven currents. Temporal trends in NSW shoreline recession have been observed during the last century (Goodwin et al. 2010), and these trends have been related to the interdecadal Pacific oscillation (IPO) (Goodwin 2005). A deeper understanding of how factors such as wave direction are related to shoreline stability is important for the development of coastal management strategies. In coastal management studies, the investigation of daily wave direction data changes over time is of particular interest, and such information can be also used in conjunction with other wave climate statistics or storm data. The data-padding and mixture-modeling approach we propose here provides a straightforward and time-efficient way to summarize and explore daily wave direction data for such purposes.

Over the last few decades, monitoring of waves along the NSW coast has been carried out through a network of Datawell BV Waverider buoys stationed at various points along the coast; one is moored off the coast of Byron Bay. The Waverider data are only available from 1977 onward; therefore, to explore wave direction over a much longer time period, we use daily mean wave direction data that have been hindcast from the 40-yr European Centre for Medium-Range Weather Forecasts (ECMWF) Re-Analysis (ERA-40) meteorological dataset. The ERA-40 (Uppala et al. 2005) is a reanalysis of meteorological observations spanning from September 1957 to August 2002. A description of the wave climate hindcasting approach used to obtain the daily mean wave direction data that we analyze in this article is given in Goodwin et al. (2010).

The dataset comprises 16 315 daily mean wave directions corresponding to dates between 1 January 1958 and 31 August 2002. The mean wave directions are recorded in degrees, which of course means that the data are circular in nature. To apply the VB approach to fit mixture models to this data, we first have to preprocess it using the padding approach that we outlined in section 2. After

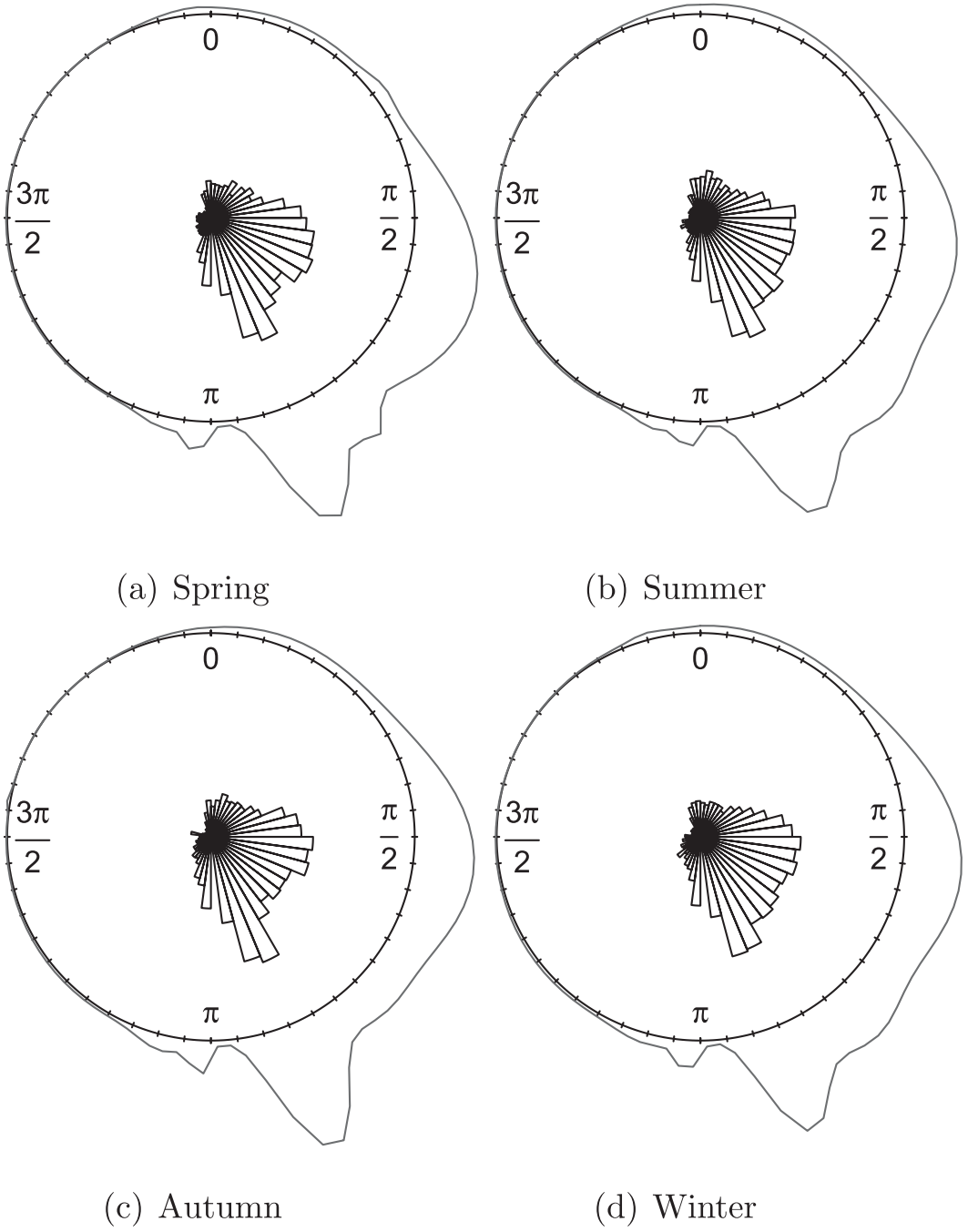


FIG. 1. Plots of the Gaussian mixture models that were fitted to the circular Byron Bay daily mean wave direction data corresponding to each season. In the center of each figure, a rose diagram of the observed data is shown.

padding, we then have a sizable dataset comprising 32 630 data points.

4. Results

We illustrate the use of VB-GMM on padded circular data by using the ideas described above to model the circular daily mean wave direction data; we fit mixture models to the data corresponding to each of the four seasons across the entire study period and also to each year of recordings. We initialized our algorithm with the number of components k set at 20 because this is a larger number than we would reasonably expect to be necessary to adequately represent this type of data. As

TABLE 1. Means μ, precisions τ, and weights w of the first, second, and third most heavily weighted components in the mixture models fitted to the circular Byron Bay daily mean wave direction data corresponding to each of the seasons. The means can lie between 0 and 2π on the circle.

	First			Second			Third		
Season	μ	$\tau^{-1/2}$	w	μ	$\tau^{-1/2}$	w	μ	$\tau^{-1/2}$	w
Spring	1.963	0.355	0.495	2.751	0.107	0.246	2.488	0.056	0.048
Summer	1.760	0.408	0.437	2.771	0.109	0.218	2.419	0.192	0.183
Autumn	1.717	0.398	0.472	2.761	0.109	0.266	2.410	0.179	0.098
Winter	1.754	0.420	0.491	2.785	0.097	0.184	2.466	0.187	0.172

described above, superfluous components can be automatically eliminated during the application of the VB algorithm.

We found in our results that the mixture models fitted to this data had between 8 and 16 components. Estimation of a suitable number of components for a mixture model is a central part of statistical inference about them; we have to estimate a suitable number of components to allow us to represent the data well. We found in this application that the top three most heavily weighted components typically represent more than 80% of the wave directions. Therefore, we concentrated on these components when interpreting results; the total actual number of components in the fit was not of particular interest to us here. In this way we are interpreting the modes of the wave directions. Often, studies examining this type of data concentrate on interpreting the overall mean of the wave direction instead. Since wave direction data are multimodal, however, the mean direction is a less informative summary statistic than the modes provided by the mixture model.

The variational approach is guaranteed to converge at least to a local minimum of the KL divergence. Carefully choosing the initialization settings for the algorithm makes it likely that the minimum found, if not the global minimum, is sufficiently close to the global minimum. The choice of initialization will depend on the model; for our application we took the approach of partitioning the data, which involved assigning the initial component membership of each observation according to which nonoverlapped equal-width interval the observed value has fallen into. We then initialized the algorithm on the basis of this partitioning, and the priors used were noninformative. This initialization is more informative than simply randomly initializing the components, and, given that weak prior information is used, it is reasonable to expect that the algorithm will lead to a suitable mixture classification of the observations in the posterior. It has been our experience in simulated data experimentation that this initialization approach works well.

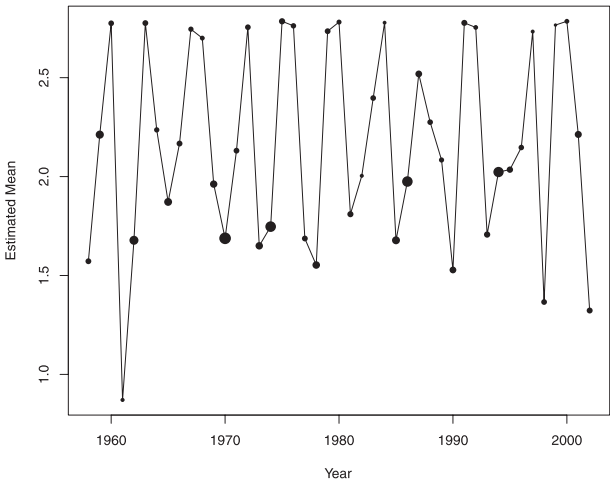


FIG. 2. The estimated mean of the most heavily weighted component of the mixture model fitted to the circular Byron Bay daily mean wave direction data for each year between 1958 and 2002. The sizes of the plotted solid circles are proportional to the sizes of the weightings of these components.

Figure 1 shows the Gaussian mixture models that were fitted for each season, providing us with a statistical estimate of the wave directions at that site. Here, we can see that, although the peak direction is fairly consistent across the seasons, some seasonal slight variation is observed. Table 1 reports the corresponding numerical values of the fitted parameters of the three most heavily weighted components for each season. Again, these suggest that the peak wave direction does not vary greatly across the summer, autumn, and winter seasons when we consider data over the period of 45 yr. The peak direction for spring appears to deviate slightly from the others; it has been noted by Goodwin et al. (2010) that the annual cycle in mean wave direction at Byron Bay is most variable between IPO phases that occur during spring, and this fact may be what is being reflected here.

For each year between 1958 and 2002, Fig. 2 plots the estimated mean of the most heavily weighted component of the mixture model fitted to the daily Byron Bay mean wave direction data for that particular year. The sizes of the plotted solid circles are proportional to the sizes of the weightings of these heaviest components. Figures 3–7 show the mixture models fitted to the data recorded for each of the years in the study period. It is interesting to see from these figures that there appears to be much variability in wave directions from year to year. This is in contrast to the seasonal fits, which were much more homogeneous. We can observe that there is a large proportion of years with similar directions, which would be useful for forming coastal protection plans. The results we have obtained appear to be consistent with previous research in suggesting that the mean wave direction

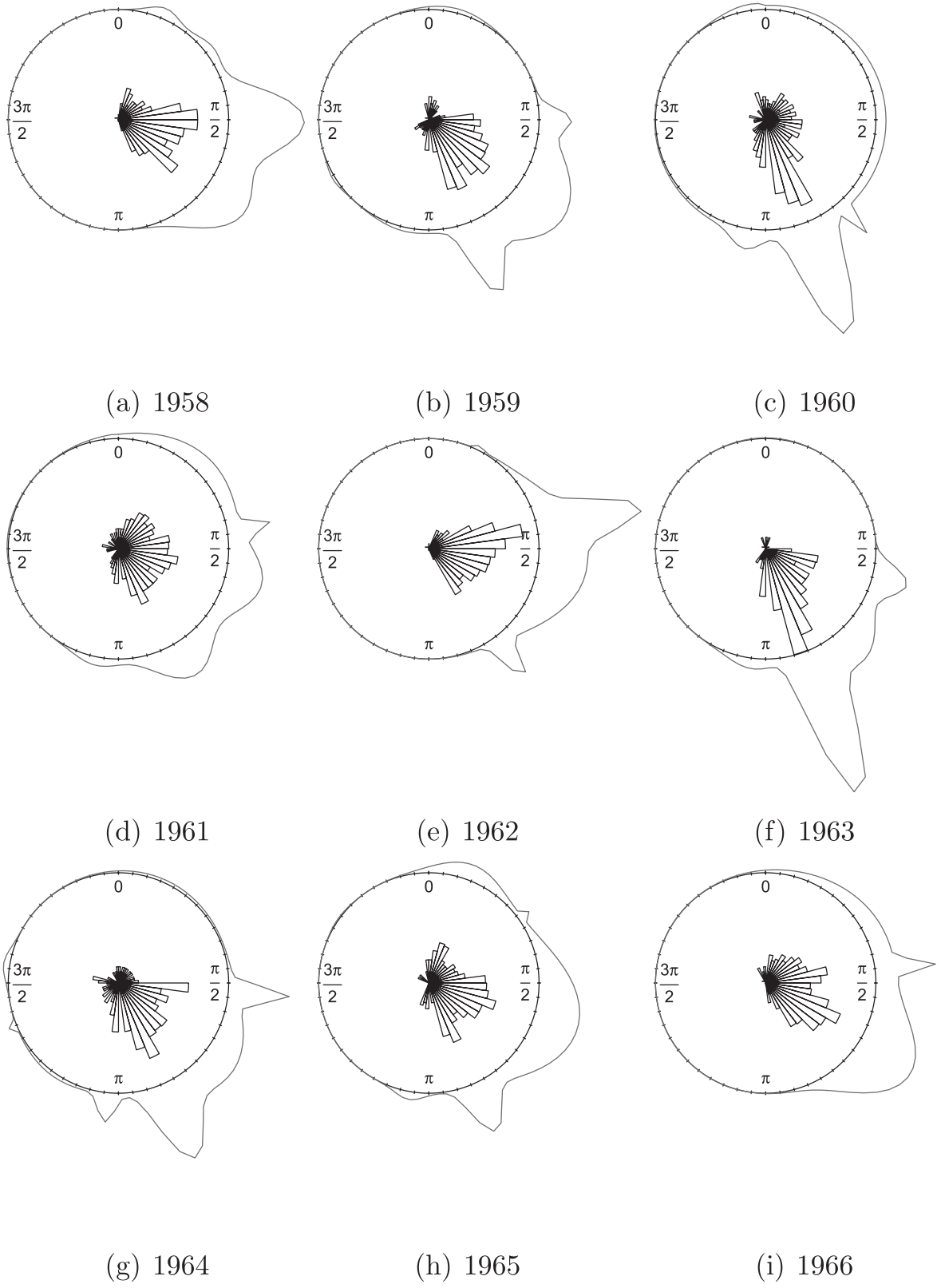


FIG. 3. Plots of the Gaussian mixture models that were fitted to the circular Byron Bay daily mean wave direction data corresponding to the years 1958–66. In the center of each figure, a rose diagram of the observed data is shown.

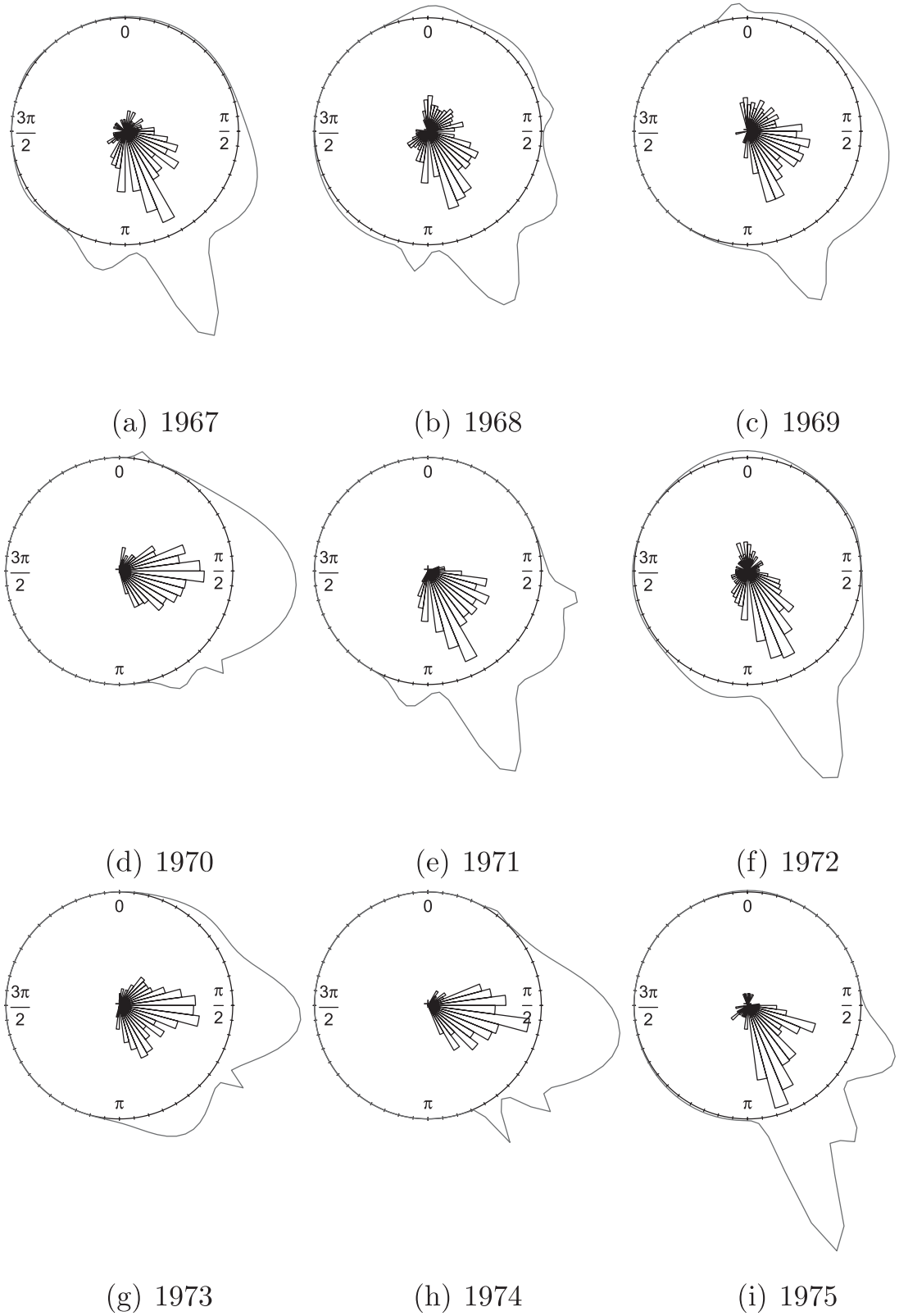


FIG. 4. As in Fig. 3, but for 1967–75.

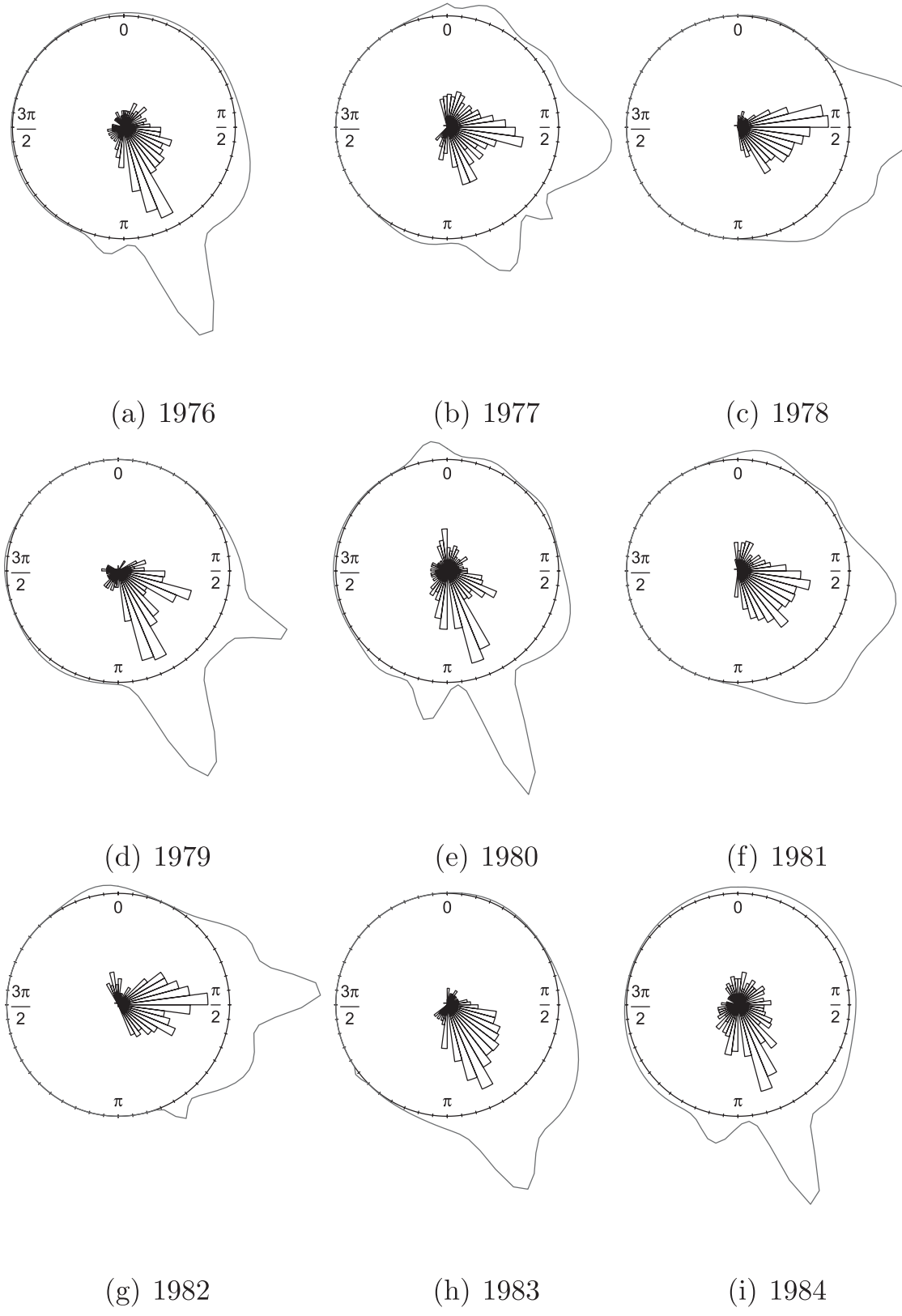
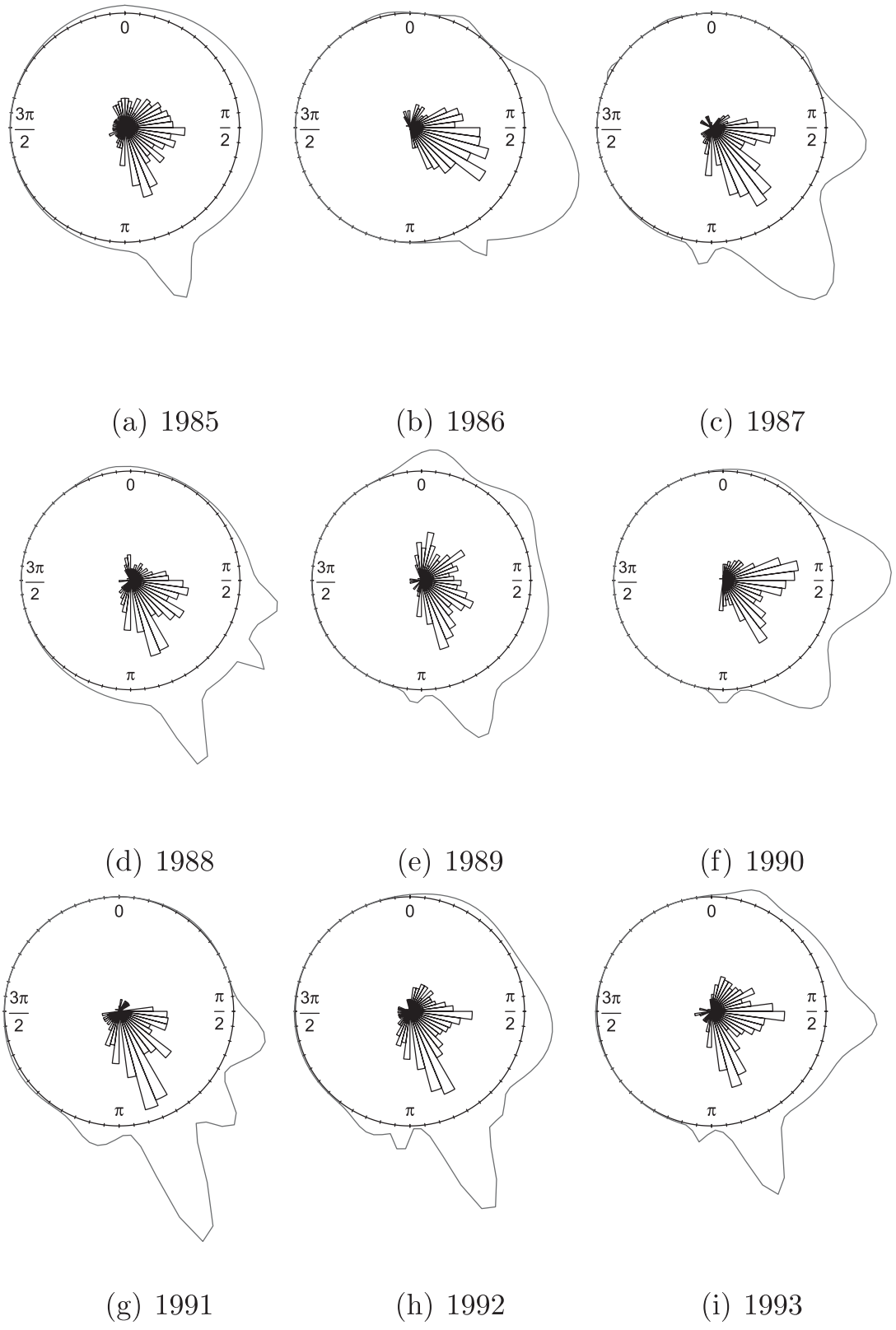


FIG. 5. As in Fig. 3, but for 1976–84.



(a) 1985 (b) 1986 (c) 1987

(d) 1988 (e) 1989 (f) 1990

(g) 1991 (h) 1992 (i) 1993

FIG. 6. As in Fig. 3, but for 1985–93.

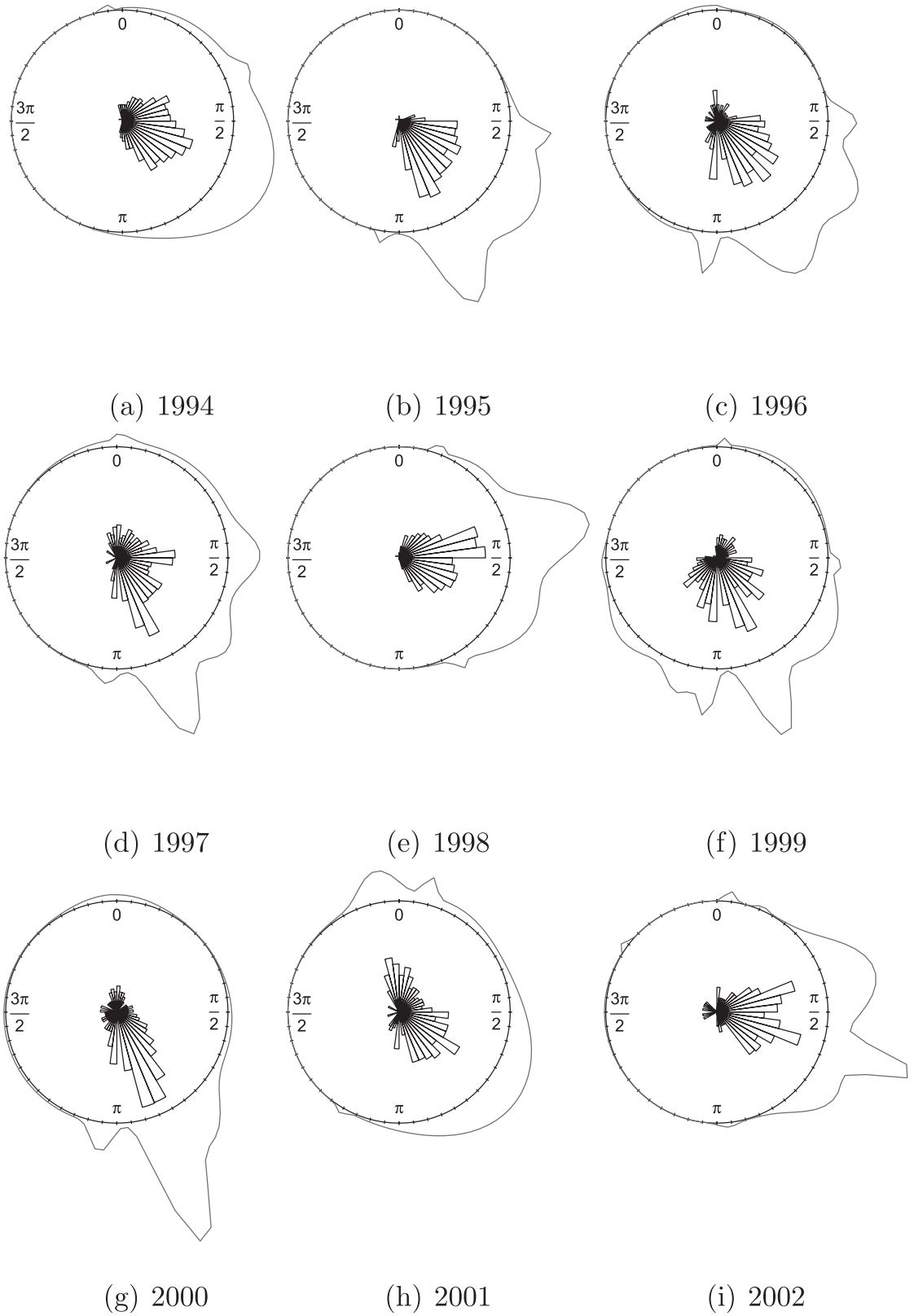


FIG. 7. As in Fig. 3, but for 1994–2002.

is highly variable over time (Goodwin 2005). Goodwin et al. (2010) state that historical analysis of shoreline change in the Byron Bay area suggests that it responds on an interannual to interdecadal period of time in phase with variability in mean wave direction; therefore, investigation of this annual variation is important for managing shoreline change.

This convenient analytical approach described here can be applied quickly, meaning that it is feasible to explore and compare various aspects of the recorded data in this way. For example, one could fit models to the wave directions observed during storm systems, or in different months, and then assess whether changes in the fitted parameters appear to be occurring. The ease of implementation means that such models can be fitted on a very regular basis to allow ongoing monitoring of variations in observations, as required. This technique will also be useful in other application areas for which circular data are observed, for example, in measuring wind direction.

Other commonly used approaches for fitting mixture models include the classical expectation–maximization (EM) algorithm approach and Bayesian MCMC-based approaches. If we were to use an EM approach for this application, it would be necessary to fit the model separately for a range of values of k and then to use a selection criterion to select the most appropriate model. Note that, dependent upon the selection criterion used, the choice of dimension for the model would vary, however. This approach would be far more time consuming than using the VB method in which both k and model parameters are estimated simultaneously. The RJMCMC algorithm of Richardson and Green (1997) can be used to simultaneously estimate k and model parameters. When implementing Bayesian MCMC-based approaches for mixture models, including RJMCMC, however, the well-known label-switching problem [see Jasra et al. (2005) for a discussion] makes inference about posterior parameters difficult. It is necessary to impose artificial identifiability constraints in the sampler to alleviate the label-switching problem when estimating the parameters, but imposing these constraints becomes increasingly challenging as the dimensionality of the model increases. Another commonly used approach is to postprocess the output from an MCMC algorithm; for example, this is the approach used in the R software package routine ''AKMix'' described in Komrek (2009). We also note that, although MCMC approaches have the attraction that in theory if the chains are run for long enough they should fully explore the posterior, in practice chains can become stuck in local minima and extremely long runs may be required to reach convergence. It can also be difficult to assess whether convergence has been reached. For these reasons, MCMC-based approaches are more time consuming to implement than the VB method for this application.

To provide an indication of the difference in implementation speed when using the VB approach described here and using the RJMCMC approach, we used AKMix to fit a mixture model via RJMCMC to the circular wave direction data that were observed in spring. There were 4004 recorded observations corresponding to the spring season in our dataset. After padding the data to account for its circular nature, the number of observations to be analyzed was 8008. For illustration we note that fitting a mixture model to the spring data using AKMix to implement RJMCMC, with 50 000 iterations and a burn-in of 10 000 iterations, took 5.5 h on a standard desktop personal computer. On top of this, further computing time would be required to postprocess the results and obtain the posterior estimates. Fitting a mixture using VB implemented in the Matlab software package took under an hour, which is significantly faster. We also note that the implementation time of the VB algorithm could be reduced further if a more efficient programming language such as C were used instead of Matlab.

5. Discussion

In this paper, we have reviewed the variational approach and shown how VB-GMM can be adapted for use in modeling circular data by taking an approach in which the data are padded at the edges. In doing so, we have proposed an effective modeling approach for circular data that can be implemented quickly and easily and that will be of particular value in settings in which there are large volumes of data to be analyzed.

We also note that we restricted our attention here to the standard VB algorithm in which components may only be eliminated and not added. With this algorithm it is possible to occasionally reach different solutions under different initialization settings, as we discussed earlier. Other types of VB algorithm have been proposed in the literature, however. For example, component-splitting VB schemes have been designed (see Wu et al. 2012 and references therein). Such schemes allow the number of mixture components to be increased as well as decreased during the convergence of the VB algorithm, thereby providing increased flexibility. This use of a component-splitting scheme could also be explored for this type of application.

Acknowledgments. We thank Professor Ian Goodwin for providing the wave direction data that were analyzed in this paper. The ERA-40 data can be obtained from the ECMWF Data Server. We are very grateful for the comments from two anonymous reviewers and the editor that have greatly improved this article.

REFERENCES

Attias, H., 1999: Inferring parameters and structure of latent variable models by variational Bayes. *Proc. 15th Conf. on Uncertainty in Artificial Intelligence,* Stockholm, Sweden, Association for Uncertainty in Artificial Intelligence, 21–30.

Bishop, C., 2006: *Pattern Recognition and Machine Learning*. Springer, 738 pp.

Celeux, G., M. Hurn, and C. P. Robert, 2000: Computational and inferential difficulties with mixture posterior distributions. *J. Amer. Stat. Assoc.,* **95,** 957–970.

Corduneanu, A., and C. M. Bishop, 2001: Variational Bayesian model selection for mixture distributions. *Proc. Eighth Int. Conf. on Artificial Intelligence and Statistics,* Key West, FL, Society for Artificial Intelligence and Statistics, 27–34.

Farrugia, P., J. Borg, and A. Micallef, 2009: On the algorithms used to compute the standard deviation of wind direction. *J. Appl. Meteor. Climatol.,* **48,** 2144–2151.

Gelman, A., J. Carlin, H. Stern, and D. Rubin, 2003: *Bayesian Data Analysis*. 2nd ed. Texts in Statistical Science, Chapman and Hall, 668 pp.

Goodwin, I., 2005: A mid-shelf, mean wave direction climatology for southeastern Australia, and its relationship to the El Niño–Southern Oscillation since 1878 A.D. *Int. J. Climatol.,* **25,** 1715–1729.

——, R. Freeman, and K. Blackmore, 2010: Decadal wave climate variability and implications for interpreting New South Wales coastal behaviour. *Proc. Australian Wind Waves Research Science Symp.,* Gold Coast, QLD, Australia, Centre for Australian Weather and Climate Research, 58–61.

Jammalamadaka, S. R., and A. Sengupta, 2001: *Topics in Circular Statistics.* World Scientific, 322 pp.

Jasra, A., C. C. Holmes, and D. A. Stephens, 2005: Markov chain Monte Carlo methods and the label switching problem in Bayesian mixture modeling. *Stat. Sci.,* **20,** 50–67.

Komrek, A., 2009: A new R package for Bayesian estimation of multivariate normal mixtures allowing for selection of the number of components and interval-censored data. *Comput. Stat. Data Anal.,* **53,** 3932–3947.

Mackay, D. J. C., 2003: *Information Theory, Inference, and Learning Algorithms*. Cambridge University Press, 628 pp.

Mahrt, L., 2011: Surface wind direction variability. *J. Appl. Meteor. Climatol.,* **50,** 144–152.

Mardia, K. V., and P. E. Jupp, 2000: *Directional Statistics*. 2nd ed. John Wiley and Sons, 429 pp.

McGrory, C. A., and D. M. Titterington, 2007: Variational approximations in Bayesian model selection for finite mixture distributions. *Comput. Stat. Data Anal.,* **51,** 5352–5367.

McLachlan, G. J., and D. Peel, 2000: *Finite Mixture Models*. John Wiley and Sons, 419 pp.

McVinish, R., and K. Mengersen, 2008: Semiparametric Bayesian circular statistics. *Comput. Stat. Data Anal.,* **52,** 4722–4730.

Richardson, S., and P. J. Green, 1997: On Bayesian analysis of mixtures with an unknown number of components (with discussion). *J. Roy. Stat. Soc.,* **59B,** 731–792.

Shand, T., and Coauthors, 2010: NSW coastal storms and extreme waves. *Proc. 19th NSW Coastal Conf.,* Batemans Bay, NSW, Australia, Australian Coastal Society, 14 pp. [Available online at http://www.coastalconference.com/2010/papers2010/Tom%20Shand%20full%20paper.pdf.]

Uppala, S. M., and Coauthors, 2005: The ERA-40 Re-Analysis. *Quart. J. Roy. Meteor. Soc.,* **131,** 2961–3012.

Weber, R., 1997: Estimators for the standard deviation of horizontal wind direction. *J. Appl. Meteor. Climatol.,* **36,** 1403–1415.

Wu, B., C. A. McGrory, and A. N. Pettitt, 2012: A new variational Bayesian algorithm with application to human mobility pattern modeling. *Stat. Comput.,* **22,** 185–203.